ADVANCES IN MATHEMATICS 10, 300–316 (1973)

Eigenvalue Approximation by the Finite Element Method

GEORGE J. FIX*

Institute for Fluid Dynamics and Applied Mathematics
University of Maryland, College Park, Maryland 20742

DEDICATED TO GARRETT BIRKHOFF

1. INTRODUCTION

By a finite element method for an eigenvalue problem $Lu = \lambda u$ involving an elliptic differential operator L, we shall mean a variational method used in conjunction with a space of piecewise polynomial functions [1]. Apparently the first mathematician to study such approximations was Courant [2] in 1943. However, it was not until 1966 in a paper by Birkhoff, de Boor, Swartz, and Wendroff [3] that a rigorous theory was developed.

In the latter, bounds are established for the error in the eigenvalues λ from which the rate of convergence as the mesh length $h > 0$ approached zero can be deduced. Such bounds have been known for many years for source problems $Lu = f$, and in fact follow readily from the best approximation properties of the variational method [1]. For eigenvalue problems, on the other hand, this matter is more delicate, and in [3] it was only after an inspired use of the *Poincaré min–max principle* that the desired bounds were obtained. Unfortunately, this work and subsequent refinements ([1, Chapter 6]–[5]) are limited only to *self-adjoint problems*—the Poincaré principle holds only in this case—and therefore does not include physically interesting nonself-adjoint problems such as the multigroup diffusion equations [6].

Based on some results due to Vainikko [7–10] the present author has developed a theory which does cover the nonself-adjoint case, and in fact assumes only that L is elliptic and has a compact inverse $T = L^{-1}$.

* This work was supported in part by the National Science Foundation under Grant NSF GP 18064, with the University of Maryland.

In this paper we give a survey of these new results, and refer the reader to the author's report [11] for further details.[1]

Perhaps the most interesting aspect of the new theory that does not occur in the self-adjoint case is the role played by the *algebraic* and *geometric multiplicities* of an eigenvalue λ_0 of L. We recall that the geometric multiplicity is the number of independent eigenfunctions associated with λ_0. The algebraic multiplicity, on the other hand, is the number of independent generalized eigenfunctions, i.e., functions u_0 satisfying $(L - \lambda_0)^l u_0 = 0$, $(L - \lambda_0)^{l-1} u_0 \neq 0$ for some $l \geqslant 1$. Our results (Section 4, Theorem 1) indicate that when these multiplicities coincide—they are always equal for self-adjoint problems—the errors in the approximate eigenvalues obtained from the finite element method are precisely the same as predicted in [3]. However, when the algebraic multiplicity is not equal to the geometric multiplicity, a loss in the rate of convergence occurs. The extent of this loss depends on the order $l_0 \geqslant 1$ of the pole in the resolvent $(z - T)^{-1}$ of the compact operator $T = L^{-1}$ at $\mu_0 = 1/\lambda_0$. This phenomenon of course has an analog in the perturbation of $N \times N$ matrices [21], and similar to the latter, it is also true that suitable averages of the eigenvalues converging to λ_0 are more accurate than the individual eigenvalues (Section 4, Theorem 2).

The approximation of eigenfunctions also involves phenomena that are not present in the self-adjoint case. For example, suppose λ_0 is an eigenvalue of L having different algebraic and geometric multiplicities. The question arises whether the finite element method approximates the eigenfunctions of λ_0 with eigenfunctions or generalized eigenfunctions. Could an approximate eigenfunction converge to a generalized eigenfunction, or conversely, a generalized eigenfunction converge to an eigenfunction? These issues are covered in Theorem 1.

In Sections 2 and 3 we outline the abstract setting for the eigenvalue problem and the finite element approximation. Our basic theorems are contained in Section 4, and the last section has examples.

2. The Eigenvalue Problem

An eigenvalue problem $Lu = \lambda u$, or more generally $Lu = \lambda Mu$ involving an elliptic differential operator L is usually formulated in the

[1] The author has recently received an unpublished manuscript by Bramble and Osborn [22] which contains an alternate theory to the one described in this paper. The final results, however, appear to be somewhat similar.

following abstract manner [12]. Let H_1, H_2 be two complex Hilbert spaces and let $a(u, v)$, $b(u, v)$ be two bounded forms on $H_1 \times H_2$; i.e.,

$$| a(u, v)| \leqslant C_1 \| u \|_{H_1} \| v \|_{H_2}, \qquad | b(u, v)| \leqslant C_1 \| u \|_{H_1} \| v \|_{H_2} \tag{2.1}$$

for all $u \in H_1$, $v \in H_2$ where C_1 is some fixed positive number. The eigenvalue problem is to find those complex numbers λ_0 (eigenvalues) such that there is a $0 \neq u_0 \in H_1$ (eigenfunction) satisfying

$$a_{\lambda_0}(u_0, v) = 0 \qquad \text{all } v \in H_2, \tag{2.2}$$

where

$$a_\lambda(u, v) = a(u, v) - \lambda b(u, v). \tag{2.2'}$$

To illustrate this let us consider the multigroup diffusion equations of reactor theory [6]. The associated differential equation is

$$-\nabla(p_i(x) \nabla u^{(i)}) + \sum_{j=1}^{i} \sigma_{ij}(x) u^{(j)} = \lambda \sum_{j=1}^{m} \sigma_j(x) u_j, \qquad 1 \leqslant i \leqslant m, \tag{2.3}$$

and holds for all x in the bounded plane region Ω. For simplicity, we use a Dirichlet condition

$$u^{(i)}(x) = 0 \qquad \text{for} \quad x \in \Gamma, \qquad 1 \leqslant i \leqslant m, \tag{2.4}$$

on the boundary Γ of Ω. The coefficients p_i, σ_{ij}, σ_j are bounded measurable functions on Ω, and typically discontinuous piecewise constant functions. As a consequence it is necessary to supplement (3)–(4) with interface conditions [6] guaranteeing the continuity of $p_i \nabla u^{(i)}$, $1 \leqslant i \leqslant m$, in Ω.

There are many variational formulations (2.2) of the multigroup equations in terms of various Hilbert spaces H_1, H_2 and forms $a(u, v)$, $b(u, v)$ [12]. However, the most common is the so-called weak form of (2.3)–(2.4). For this formulation we take

$$H = H_1 = H_2 = \underbrace{H_0^1(\Omega) \times \cdots \times H_0^1(\Omega)}_{m \text{ times}}, \tag{2.5}$$

where $H_0^1(\Omega)$ is the Sobolev space [13] of functions which vanish on Γ and whose first derivatives are square integrable in Ω. The associated norm of a function $u = (u^{(1)}, \ldots, u^{(m)})$ in H is

$$\| u \|_H = \left\{ \sum_{i=1}^{m} \int_\Omega [| \nabla u^{(i)} |^2 + | u^{(i)} |^2] \, dx \right\}^{1/2}, \tag{2.6}$$

and the bilinear forms are

$$a(u, v) = \sum_{i=1}^{m} \int_{\Omega} \left[p_i \nabla u^{(i)} \cdot \nabla v^{(i)} + \sum_{j=1}^{i} \sigma_{ij} u^{(j)} v^{(i)} \right] dx, \tag{2.7}$$

$$b(u, v) = \sum_{i,j=1}^{m} \int_{\Omega} \sigma_j u^{(j)} v^{(i)}. \tag{2.8}$$

The variational Eq. (2.2) with a and b given by (2.7)–(2.8) is formally obtained from (2.3) by multiplying the latter by an arbitrary $v^{(i)} \in H_0^1(\Omega)$, integrating by parts using the boundary and interface conditions, and then summing over i, $1 \leqslant i \leqslant m$. This incidentally shows that if the boundaries and interfaces are smooth enough to permit these operations—this is never a problem in practice since they are usually piecewise analytic—then any eigenvalue $\lambda = \lambda_0$ of the "classical formulation" (2.3)–(2.4) is also an eigenvalue of (2.2). The converse, however, is not true in general. It is possible that (2.1) has an eigenvalue λ_0 and eigenfunction u_0 whose second derivatives do not exist, and hence a u_0 for which we cannot reverse the above steps to deduce (2.2)–(2.3) from (2.1). This latter phenomenon never occurs in reactor theory, on the other hand, since the coefficients p_i that arise from such problems are positive in Ω. This implies that (2.3) is elliptic [13] and therefore has solutions that are smooth in Ω [13], except possibly on certain sets of measure zero [14].

Conditions analogous to "ellipticity" are also needed in the abstract setting, since the eigenvalue problem (2.2), where a and b are only known to be bounded forms on $H_1 \times H_2$, is not very interesting. It is possible that no eigenvalues exist, or at the other extreme, every number λ_0 is an eigenvalue [11]. To state such "ellipticity" conditions, we first note that (2.1) and the Riesz Representation Theorem permit us to write

$$a(u, v) = (Au, v)_{H_2}, \qquad b(u, v) = (Bu, v)_{H_2}, \tag{2.9}$$

where A and B are bounded transformations from H_1 to H_2. Our first assumption is that A has a bounded inverse A^{-1} from H_2 to H_1, an assumption which is usually written

$$\inf_{\|u\|_{H_1}=1} \sup_{\|v\|_{H_2}=1} |a(u, v)| \geqslant C_2, \qquad \inf_{\|v\|_{H_2}=1} \sup_{\|u\|_{H_1}=1} |a(u, v)| \geqslant C_2 \tag{2.10}$$

for some positive number C_2 (see [12] and [15]). For the multigroup equations (2.10) holds provided the coefficients p_i are positive in $\bar{\Omega}$, i.e. (2.3) is elliptic.

Our second assumption is much more substantial, and in particular, we shall assume that

$$T = A^{-1}B \tag{2.11}$$

is a compact operator on H_1. This assumption is valid for the multigroup equations [16] and for most but not all elliptic problems [13]. In any case, the theory which we shall describe in this paper depends fundamentally on this property of T.

To see the implications of this assumption for the eigenvalue problem (2.2), we note that because of (2.10), (2.2) is equivalent to

$$\lambda_0 T u_0 = u_0 ; \tag{2.12}$$

i.e., the reciprocals of the nonzero eigenvalues of T are precisely the eigenvalues of (2.2). We recall that since T is compact, its eigenvalues form a countable infinite subset of the complex plane having a limit point only at the origin. Each eigenvalue $\mu_0 = 1/\lambda_0$ is a pole of order $l_0 \geqslant 1$ of the resolvent

$$R(z, T) = (z - T)^{-1}, \tag{2.13}$$

and the integer l_0 is often called the *index of* λ_0. If $\mu = 1/\lambda \neq 0$ is not an eigenvalue of T, then $R(\mu, T)$ is a bounded operator on H_1. This means that the source problem $(I - \lambda T)u = f$, or equivalently

$$a_\lambda(u, v) = a(f, v) \qquad \text{all } v \in H_2 \,^2 \tag{2.14}$$

can be solved uniquely for u given any $f \in H_1$.

To describe the associated eigenfunctions and *generalized eigenfunctions* associated with the nonzero eigenvalue $\mu_0 = 1/\lambda_0$ of T, we set

$$Q = \frac{1}{2\pi i} \int_{\Gamma_0} R(z, T)\, dz, \tag{2.15}$$

[2] Because of (2.10) any bounded functional F on H_2 can be written $F(v) = a(f, v)$ for an appropriate $f \in H_1$. Hence (2.14) is equivalent to the more usual form ([12], [15])

$$a_\lambda(u, v) = F(v) \qquad \text{all} \quad v \in H_2$$

of the source problem.

where Γ_0 is a circle about μ_0 containing no other eigenvalues of T. Then [17] Q is a bounded projection of H_1 onto the finite dimensional subspace $N(\lambda_0, l_0)$ consisting of generalized eigenfunctions of order at most l_0. More precisely, each $u_0 \in N(\lambda_0, l_0)$ satisfies $(I - \lambda_0 T)^{l_0} u_0 = 0$, or what is the same,

$$a_{\lambda_0}(u^{(i)}, v) = a(u^{(i+1)}, v) \qquad \text{all } v \in H_2 \tag{2.16}$$

for some sequence $u^{(1)} = u_0$, $u^{(2)}, \ldots, u^{(l_0)}$ in H_1 where $u^{(l_0+1)} = 0$. The dimension of $N(\lambda_0, l_0)$ is called the *algebraic multiplicity* of λ_0. More generally, we set

$$N(\lambda_0, j) = \{u_0 \in H_1 \mid (I - \lambda_0 T)^j u_0 = 0\}. \tag{2.17}$$

Then $N(\lambda_0, j) = N(\lambda_0, l_0)$ for all $j \geqslant l_0$ and $N(\lambda_0, 1)$ is the space of eigenfunctions of (2.2). The dimension of $N(\lambda_0, 1)$ is called the *geometric multiplicity* of λ_0.

The *transpose problem*

$$a_{\lambda_0}(u, v_0) = 0 \qquad \text{all } u \in H_1 \tag{2.18}$$

is also important to our theory. It is not difficult to show that (2.2) and (2.18) are equivalent in the sense that they have the same eigenvalues [11]. In fact, the eigenvalues of (2.2) are precisely conjugates of the eigenvalues of the compact operator T_* defined by

$$a(u, T_* v) = a(Tu, v) \qquad \text{all } u \in H_1, \quad v \in H_2.$$

In addition, $1/\lambda_0$ is a pole of order l_0 of $(z - T_*)^{-1}$ if and only if it is a pole of order l_0 of $(z - T)^{-1}$, and the dimension of generalized eigenfunction space

$$N_*(\lambda_0, l_0) = \{v_0 \in H_2 \mid (I - \lambda_0 T_*)^{l_0} v_0 = 0\}$$

of the transposed problem is equal to the dimension of $N(\lambda_0, l_0)$.

In the *self-adjoint case* we have

$$H = H_1 = H_2,$$

and

$$a(u, v) = a(v, u), \qquad b(u, v) = b(v, u)$$

for all $u, v \in H$. Thus, $T = T_*$ and the problems (2.2), (2.18) are identical. Moreover, every nonzero eigenvalue $1/\lambda_0$ of T is a first order pole ($l_0 = 1$) of the resolvent $(z - T)^{-1}$, and hence $N(\lambda_0, j) = N(\lambda_0, 1)$ for all $j \geqslant 1$—every generalized eigenfunction is a true eigenfunction. In particular, the algebraic and geometric multiplicities coincide.

3. The Finite Element Approximation

To approximate the eigenvalues and eigenfunctions of (2.2) we introduce a family of finite dimensional subspaces $S_i^h \subset H_i$, $i = 1, 2$, depending on a parameter $0 < h < 1$ going to zero. The idea is to solve (2.2) on these subspaces; i.e., we seek those numbers λ_h such that there is a $0 \neq u_h$ in S_1^h satisfying

$$a(u_h, v^h) = \lambda_h b(u_h, v^h) \qquad \text{all } v^h \in S_2^h. \tag{3.1}$$

To accurately model the type of setting that arises from the finite element method we introduce two assumptions. The first concerns the invertibility of the form $a(u, v)$ on $S_1^h \times S_2^h$. In particular, we assume that the "ellipticity conditions" (2.10) hold with H_i replaced with S_i^h; i.e.,

$$\begin{aligned} \inf_{\|u\|^h_{H_1}=1} \sup_{\|v^h\|_{H_2}=1} |a(u^h, v^h)| &\geqslant C_2, \\ \inf_{\|v^h\|=1} \sup_{\|u^h\|=1} |a(u^h, v^h)| &\geqslant C_2, \end{aligned} \tag{3.2}$$

where the inf and sup are taken for $u^h \in S_1^h$, and $v^h \in S_2^h$ (see [12, 15]).

To see the implications of (3.2) let us select a basis $\phi_1^h, \ldots, \phi_N^h$ for S_1^h and $\psi_1^h, \ldots, \psi_N^h$ for S_2^h. Thus the test function $u_h \in S_1^h$ in (3.1) can be written

$$u_h = \sum_{j=1}^{N} q_j^h \phi_j^h, \tag{3.3}$$

while all trial functions $v^h \in S_2^h$ have the form

$$v^h = \sum_{j=1}^{N} p_j^h \psi_j^h. \tag{3.4}$$

In particular, (3.1) reduces to the matrix eigenvalue problem

$$A^h q^h = \lambda_h B^h q^h \tag{3.5}$$

in the weights q^h of u_h, where the (i, j) entries of the $N \times N$ matrices A^h, B^h are $a(\phi_i{}^h, \psi_j{}^h)$, $b(\phi_i{}^h, \psi_j{}^h)$ respectively [11]. The inequalities (3.2) are easily seen to be equivalent to the invertibility of A^h [11], hence with the former we can be assured that (3.1) has N isolated eigenvalues λ_h. These, incidentally, are also the eigenvalues of the transposed problem

$$a(u^h, v_h) = \lambda_h b(u^h, v_h) \qquad \text{all } u^h \in S_1{}^h. \tag{3.6}$$

It is possible to formulate (3.1) in terms of an operator T^h on the Hilbert space H_1. To see this let us introduce the operator P^h on H_1 satisfying

$$a(P^h u, v^h) = a(u, v^h) \qquad \text{all } v^h \in S_2{}^h. \tag{3.7}$$

Such an operator exists and is a projection of H_1 onto $S_1{}^h$. It is constructed by writing

$$P^h u^h = \sum_{j=1}^{N} q_j{}^h \phi_j{}^h,$$

and computing the weights q^h from

$$A^h q^h = f^h,$$

where f^h is the $N \times 1$ vector $[a(u, \psi_j{}^h)]_{j=1}^N$. Our desired operator T^h is the composition of P^h with T, i.e.,

$$T^h = P^h T,$$

and (3.1) is equivalent to

$$\lambda_h T^h u_h = u_h .$$

One of the most important properties of finite element spaces $S_1{}^h$, $S_2{}^h$ is that T^h closely approximates T for sufficiently small $h > 0$ [11]. We introduce this property into the abstract setting at this point by formally assuming

$$\| T - P^h T \| = \{ \sup_{\|u\|_{H_1}=1} \|(T - P^h T)u\|_{H_1} \} \to 0 \qquad \text{as} \quad h \to 0. \tag{3.8}$$

Since generalized eigenfunctions $u_0 \in N(\lambda_0, l_0)$ are in the range of T, it follows that they can be approximated by elements in S_1^h; i.e.,

$$\epsilon^h = \sup_{u_0 \in N(\lambda_0, l_0)} \inf_{w^h \in S_1^h} (\| u_0 - w^h \|_{H_1} / | u_0 |_{H_1}) \to 0 \qquad \text{as} \quad h \to 0, \tag{3.9}$$

and similarly

$$\epsilon_*^h = \sup_{v_0 \in N(\lambda_0, l_0)} \inf_{w_*^h \in S_2^h} (\| v_0 - w_*^h \|_{H_2} / \| v_0 \|_{H_2}) \to 0 \qquad \text{as} \quad h \to 0 \tag{3.10}$$

for the transpose problem.

The function $P^h u_0 \in S_1^h$ is of course not the best approximation w_0^h to the generalized eigenfunction $u_0 \in N(\lambda_0, l_0)$ in S_1^h, but it is nevertheless a very good one. Indeed, we note that the projection P^h is bounded,

$$\| P^h \| \leqslant C_1/C_2 ,$$

where C_1 is the constant in (2.1) and C_2 is the constant in (3.2). Thus

$$\begin{aligned} \|(I - P^h)u_0\|_{H_1} &= \|(I - P^h)(u_0 - w_0^h)\|_{H_1} \\ &\leqslant (1 + \| P^h \|) \| u_0 - w_0^h \|_{H_1} \leqslant [1 + (C_1/C_2)]\, \epsilon^h; \end{aligned} \tag{3.11}$$

i.e., the error in the approximation $P^h u_0$ to u_0 differs from the error in the best approximation w_0^h by a constant multiple independent of h. We shall make repeated use of this fact in the next section.

4. Error Estimates

Let $\mu_0 = 1/\lambda_0$ be an eigenvalue of T, and let $l_0 \geqslant 1$ be the index of λ_0. Since $\| T - P^h T \| \to 0$ as $h \to 0$ and T is compact, it is not difficult to show [11, 20] that there is an eigenvalue λ_h and eigenfunction u_h of (3.1) such that λ_h converges to λ_0 and u_h converges to some eigenfunction u_0 as $h \to 0$. Our goal is to take the next step and relate these errors to approximability properties of the spaces S_i^h $(i = 1, 2)$ as measured by ϵ^h, ϵ_*^h in (3.9)–(3.10). We shall also want to consider the errors in the generalized eigenfunctions, and for simplicity we let

$N_h(\lambda_h, j)$ denote the generalized eigenfunctions of order j associated with λ_h; i.e.,

$$N_h(\lambda_h, j) = \{u_h \in S_1^h \mid (I - \lambda_h T^h)^j u_h = 0\},$$

where $T^h = P^h T$.

THEOREM 1. *Let* (2.1), (2.10), (3.2), (3.8) *hold, and let* T *be a compact operator on* H_1. *Then there is an eigenvalue* λ_h *of* (3.1) *such that*

$$|\lambda_h - \lambda_0|^{l_0} \leqslant C\epsilon^h \epsilon_*^h. \tag{4.1}$$

If $u_h \in N_h(\lambda_h, j)$ *for any* $j \geqslant 1$ *and* $\|u_h\|_{H_1} = 1$, *then*

$$\inf_{u_0 \in N(\lambda_0, l_0)} \|u_h - u_0\|_{H_1} \leqslant C\epsilon^h. \tag{4.2}$$

In addition, if $1 \leqslant j \leqslant k \leqslant l_0$, *then*

$$\inf_{u_0 \in N(\lambda_0, k)} \|u_h - u_0\|_{H_1} \leqslant C(\epsilon^h + |\lambda_h - \lambda_0|^{k+1-j}). \tag{4.3}$$

In (4.1)–(4.3) C *is a positive constant which depends only on* T, λ_0, *and the families* $\{S_i^h\}_{0<h<1}$ *but not on* $h > 0$.

THEOREM 2. *Let the assumptions of Theorem* 1 *hold. Denote by* $\{\lambda_h^{(i)} \mid 1 \leqslant i \leqslant m\}$ *the set of eigenvalues repeated according to algebraic multiplicity which converge to* λ_0 *as* $h \to 0$. *Then*

$$\left| \lambda_0^{-1} - \left[(1/m) \sum_{j=1}^{m} (\lambda_h^{(j)})^{-1} \right] \right| \leqslant C\epsilon^h \epsilon_*^h, \tag{4.4}$$

where C *is the same constant that appears in Theorem* 1.[3]

In the self-adjoint case when $S^h = S_1^h = S_1^{h}$, (4.1) reduces to the estimate proved in [2]; namely the error $\lambda_h - \lambda_0$ is of order

$$\inf_{w^h \in S^h} |a(u_0 - w^h, u_0 - w^h)|. \tag{4.5}$$

However, even in this case (4.1) offers some improvement over the earlier estimates. In the latter, the bound for $\lambda_h - \lambda_0$ involved not only (4.5) but also the error $\lambda_h' - \lambda_0'$ for eigenvalues λ_0' whose magnitude is smaller than λ_0. Our new estimate (4.1) removes the dependence

[3] This result is due to Bramble and Osborn [22].

on the latter, a fact which has some importance in problems like the multigroup diffusion equations (see Section 5).

Observe that as the order l_0 of the pole of the resolvent $(z - T)^{-1}$ at $\mu_0 = 1/\lambda_0$ increases, the accuracy in the approximate eigenvalue λ_h decreases. That such a loss does in fact occur is shown in [11] with examples on sequence spaces where λ_0 and λ_h can be computed explicitly. Theorem 2 states, on the other hand, that no loss is experienced even for large l_0 if we use the harmonic average

$$m\left[\sum_{j=1}^{m} \lambda_h^{(j)-1}\right]^{-1}$$

to approximate λ_0. However, from a practical point of view this result is not as important as it might appear. For one thing, $l_0 = 1$ for most physically interesting eigenvalues. Second, in actual computation it may not always be clear which of the eigenvalues λ_h to use in the average, especially if (2.1) has one or more eigenvalues λ_0' close to λ_0.

The eigenfunction estimates state that if u_h is a generalized eigenfunction of order $j \geqslant 1$ of T^h, then there is a generalized eigenfunction u_0 of T having maximal order l_0 which differs from u_h by $O(\epsilon^h)$. It is true that u_h also differs from a generalized eigenfunction of lower order k, provided $k \geqslant j$; however, this error is in general larger than the previous one. For example, if u_h is an eigenfunction of T^h, then it differs from an eigenfunction of T by the amount $O(\epsilon^h) + O(\epsilon^h \epsilon_*{}^h)^{1/l_0}$. The latter term dominates if $l_0 > 2$.

We now briefly sketch the proof of Theorem 1 for the case $l_0 = 1$. The starting point is the identity

$$a_{\lambda_0}(u, v) = a_{\lambda_h}(u, v) + (\lambda_h - \lambda_0)\, b(u, v) \tag{4.5}$$

which is valid for all $u \in H_1$ and $v \in H_2$. To derive an expression for eigenvalue error $\lambda_h - \lambda_0$ we set $u = u_h$—the eigenfunction of (3.1) associated with λ_h—and $v = v_0 \in N_*(\lambda_0, 1)$. Then the term on the left drops out since $a_{\lambda_0}(u, v_0) = 0$ for all $u \in H_1$, and we are left with

$$(\lambda_0 - \lambda_h)\, b(u_h, v_0) = a_{\lambda_h}(u_h, v_0).$$

As was mentioned at the beginning of the section, u_h converges to some eigenfunction $u_0 \in N(\lambda_0, 1)$. Moreover, we may choose $v_0 \in N(\lambda_0, 1)$ such that $b(u_0, v_0) = 1$ [11]. Hence

$$|\lambda_0 - \lambda_h| \leqslant C\, |a_{\lambda_h}(u_h, v_0)|. \tag{4.6}$$

In addition, $a_{\lambda_h}(u_h , w_*{}^h) = 0$ all $w_*{}^h \in S_2{}^h$ by (3.1), thus

$$| \lambda_0 - \lambda_h | \leqslant C \inf_{w_*{}^h \in S_2{}^h} | a_{\lambda_h}(u_h , v_0 - w_*{}^h)|. \tag{4.7}$$

Before estimating the right side of (4.7) let us return to (4.5) and get a similar expression for the eigenfunction error $u_0 - u_h$. Putting $u = u_h$ and leaving $v \in H_2$ arbitrary in (4.5) gives

$$a_{\lambda_0}(u_h , v) = a_{\lambda_h}(u_h , v) + (\lambda_h - \lambda_0)\, b(u_h , v).$$

We now use the projection Q onto $N(\lambda_0 , l_0) = N(\lambda_0 , 1)$ to write

$$a_{\lambda_0}(u_h , v) = a_{\lambda_0}((I - Q)u_h , v).$$

The form a_{λ_0} is invertible on the null space of Q, i.e.,

$$C \sup_{\|v\|_{H_2}=1} | a_{\lambda_0}((I - Q)u, v)| \geqslant \|(I - Q)u \|_{H_1}$$

(see [11]). Thus with (4.7) we conclude

$$\|(I - Q)u_h \|_{H_1} \leqslant C \sup_{\|v\|_{H_2}=1} | a_{\lambda_h}(u_h , v)|. \tag{4.8}$$

Observe that the right sides of (4.7) and (4.8) both involve $a_{\lambda_h}(u_h , v)$ where in the first case $v = v_0 - w_*{}^h$ and in the second case it is arbitrary. To estimate such terms we use $\lambda_h P^h T u_h = u_h$ to write

$$a_{\lambda_h}(u_h , v) = \lambda_h a([P^h T - T]u_h , v).$$

We now factor u_h into its component Qu_h in $N(\lambda_0 , l_0) = N(\lambda_0 , 1)$ and its component $(I - Q)u_h$ in the null space of Q. Recalling that $\lambda_0 TQ = Q$, this gives

$$a_{\lambda_h}(u_h , v) = (\lambda_h/\lambda_0)\, a((I - P^h)\, Qu_h , v) + \lambda_h a((I - P^h)\, T(I - Q)u_h , v).$$

Now, since $Qu_h \in N(\lambda_0 , l_0)$ we have

$$\|(I - P^h)\, Qu_h \| \leqslant C\epsilon^h$$

from (3.11). Thus using (2.1)

$$| a_{\lambda_h}(u_h , v)| \leqslant C \| v \|_{H_2}\{\epsilon^h + \| T - P^h T \| \|(I - Q)u_h \|_{H_1}\}. \tag{4.10}$$

In is now a simple matter to complete the proof with (2.10). In particular, substituting it into (4.8) and using the fact that $\| T - P^h T \| \to 0$ as $h \to 0$, we obtain

$$\|(I - Q)u_h\|_{H_1} \leqslant C\epsilon^h,$$

which is the same as (4.2). On the other hand, setting $v = v_0 - w_*{}^h$ in (4.10) and substituting the result into (4.7), we obtain

$$|\lambda_h - \lambda_0| \leqslant C\epsilon_*{}^h\{\epsilon^h + \| T - P^h T \| \|(I - Q)u_h\|_{H_1}\} \leqslant C\epsilon_*{}^h\epsilon^h.$$

5. Examples

Let us return to the multigroup diffusion equations (2.3), verify that the assumptions introduced in Section 3 hold for the usual finite element approximations, and then interpret the conclusions of Theorem 1 for this problem. It is convenient to distinguish between the case where there are interfaces, and the case where the coefficients P_i, σ_{ij}, σ_i are smooth.

We start with the latter and assume at the same time that Ω has a C^∞ boundary Γ. It follows [13]—assuming the same is true of the coefficients—that each component $u_0^{(i)}$ of the eigenfunctions $u_0 = (u_0^{(i)},\ldots, u_0^{(m)})$ is in $C^\infty(\bar{\Omega})$. To construct a finite element approximation we subdivide Ω into triangles (or rectangles) with curvilinear elements on the boundary [3],[4] and let $h \geqslant 0$ denote the maximum diameter of an element. We let $S_1{}^h = S_2{}^h$ be equal to

$$\frac{S^h \times \cdots \times S^h}{m \text{ times}},$$

where S^h consists of piecewise polynomials of degree $k - 1$ in each subdivision except of course for the curved elements near the boundary where they have a slightly more complicated form [3]. For $S_1{}^h \subset H_1$ to be true, we of course need to require that each function in $S_1{}^h$ vanishes on Γ.

[4] An interesting and practical alternative to the curved or isoparametric elements for regions Ω having curved boundaries is provided by Lagrange multipliers [12, 18]. This requires a different variational method, i.e., different spaces H_1, H_2 and forms a, b than (2.5), (2.7)–(2.8). However, such a setting is covered by our theory and the reader is referred to [11] for details.

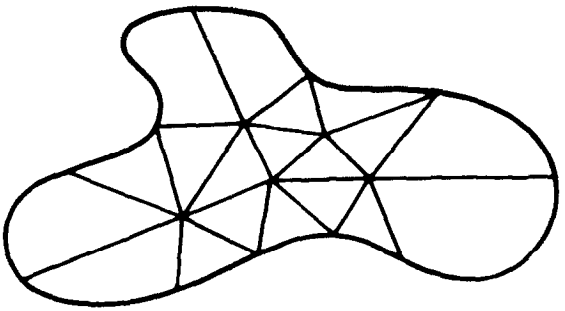

FIG. 1. The region Ω.

That the form $a(u, v)$ defined by (2.7) is invertible on $S_1^h \times S_2^h$, i.e., (3.2) holds, can be verified by a simple but lengthy calculation. This for example is done in [12]. Turning to the approximability assumptions (3.8)–(3.10), we recall [3] that if $w \in H^k(\Omega) \cap H_0^1(\Omega)$ so that

$$\| u \|_{H^k(\Omega)} = \left\{ \sum_{0 \leq \alpha_1 + \alpha_2 \leq k} \int_\Omega \left| \frac{\partial^{\alpha_1+\alpha_2} w(x)}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}} \right|^2 dx \right\}^{1/2} < \infty,$$

then there is a $w_h \in S^h$ satisfying

$$\| w - \tilde{w}_h \|_{H^l(\Omega)} \leq Ch^{k-l} \| u \|_{H^k(\Omega)} \tag{5.1}$$

for $l = 0$ and 1. In particular, if each component $u^{(i)}$ of $u = (u^{(i)}, \ldots, u^{(m)})$ is in $H^k(\Omega) \cap H_0^1(\Omega)$, then

$$\| u - P^h u \|_{H_1} \leq Ch^{k-1} \sum_{i=1}^m \| u^{(i)} \|_{H^k(\Omega)}. \tag{5.2}$$

[see (3.11)]. For any given $u \in H_1$, each component of Tu will be in $H^3(\Omega) \cap H_0^1(\Omega)$ [13], hence

$$\| (I - P^h) Tu \|_H \leq Ch^2 \sum_{i=1}^m \| (Tu)^{(i)} \|_{H^3(\Omega)}.$$

Since the norm on the right is equivalent to $\| u \|_{H_1}$ [13], we have

$$\| T - P^h T \| = O(h^2). \tag{5.3}$$

The errors ϵ^h and $\epsilon_*{}^h$ are much smaller than (5.3). To fix ideas, let us consider only the unique positive eigenvalue $\mu_0 = 1/\lambda_0$ of T having largest modulus (see Section 2). The algebraic and geometric multiplicities coincide for this eigenvalue, hence $l_0 = 1$ and $N(\lambda_0, l_0)$, $N_*(\lambda_0, l_0)$ are multiples of the eigenfunctions u_0, v_0. Both of these eigenfunctions are in $C^\infty(\bar{\Omega})$, hence from (5.1) we obtain

$$\epsilon^h = O(h^{k-1}), \qquad \epsilon_*{}^h = O(h^{k-1}). \tag{5.4}$$

In particular, Theorem 1 states that the finite element method produces a simple eigenvalue λ_h such that

$$(\lambda_h - \lambda_0) = O(h^{2k-2}), \tag{5.5}$$

and the associated eigenfunction $u_h = (u_h^{(i)}, \ldots, u_h^{(m)})$ has the error (the error in the L_2 norms is $O(h^k)$ (see [11])

$$\| u_h - u_0 \|_{H_1} = \left\{ \sum_{j=1}^{m} \int_\Omega | \nabla(u_h^{(j)} - u_0^{(j)})|^2 + | u_h^{(j)} - u_0^{(j)} |^2 \right\}^{1/2} = O(h^{k-1}). \tag{5.6}$$

Similar estimates are valid for the eigenfunctions v_0, v_h of the transposed problem.

When the coefficients p_j are discontinuous in Ω, the situation is somewhat more complicated since certain of the eigenfunctions may have singular derivatives. To fix ideas suppose Ω is a square with interfaces shown in Fig. 2. Thus the coefficients p_j, σ_{ij}, σ_i are constant in Ω_0 and Ω, and discontinuous across the boundary of Ω_0. Physically, Ω_0 represents the core of the reactor and Ω_1 is the reflector [6]. It is known [12] that the dominant eigenfunction u_0 associated with the largest positive eigenvalue $\mu_0 = 1/\lambda_0$ of T has unbounded first derivatives at A, B, C, and D. Each component $u_0^{(i)}$ of u_0 is only in the fractional Sobolev space $H^{1+\mu}(\Omega) \cap H_0{}^1(\Omega)$ [13] for some $0 < \mu < 1$ depending on

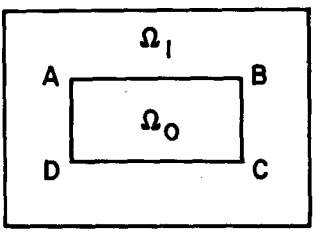

FIG. 2. The region Ω with interfaces.

the coefficients p_j. Thus if we use the usual finite element approximation,

$$\epsilon^h = O(h^\mu), \qquad \epsilon_*{}^h = O(h^\mu), \tag{5.7}$$

and the eigenvalue error is only $O(h^{2\mu})$ independent of the degree $k - 1$ of the piecewise polynomial functions used. However, not all of the eigenvalues of the multigroup problem have singular eigenfunctions. Some in fact have continuous derivatives of all orders [14]. For the latter the eigenvalue errors are precisely (5.5); i.e., there is no "pollution" due to less accurate approximation of other eigenvalues having singular eigenfunctions.

By appropriate mesh refinement we can recapture the eigenvalue error (5.5) for the eigenvalues $\mu_0 = 1/\lambda_0$ having singular eigenfunctions. Since mesh refinement has the effect of reducing ϵ^h, $\epsilon_*{}^h$ from (5.7) to (5.4). An alternate scheme is to use functions which match the singularity of u_0, v_0 at the corners A, B, C, D [19]. This interestingly enough leads to *different test and trial functions*, i.e., $S_1{}^h \neq S_2{}^h$. Specifically, we let

$$S_1{}^h = S^h \times \cdots \times S^h,$$

where S^h is the span of the finite elements and the singular functions associated with u_0. Since u_0 minus a suitable linear combination of these singular functions is smooth,

$$\epsilon^h = O(h^{k-1}).$$

Similarly, we take

$$S_2{}^h = S_*{}^h \times \cdots \times S_*{}^h,$$

where $S_*{}^h$ is the span of the finite elements plus the singular functions associated with eigenfunction v_0 of the transposed problem. Thus

$$\epsilon_*{}^h = O(h^{k-1})$$

and the errors in λ_0, u_0 are given by (5.5)–(5.6).

References

1. G. Strang and G. Fix, "An Analysis of the Finite Element Method," Prentice–Hall, New York (to appear).
2. R. Courant, *Bull. Amer. Math. Soc.* **40** (1943), 1–23.
3. G. Birkhoff, C. deBoor, B. Swartz, and B. Wendroff, *SIAM J. Numer. Anal.* **13** (1966), 188–203.

4. P. G. CIARLET, M. H. SCHULTZ, AND R. S. VARGA, *Numer. Math.* **12** (1968), 120–133.
5. J. G. PIERCE AND R. S. VARGA, *SIAM J. Numer. Anal.* **9** (1972), 137–151.
6. S. GLASSTONE AND M. C. EDLUND, "The Elements of Nuclear Reactor Theory," Von Nostrand, New York, 1952.
7. G. M. VIANIKKO, *USSR Comput. Math. and Math. Phys.* **7** (1967), 18–32.
8. G. M. VIANIKKO, *Sh. vyschisl. Mat. mat. Fiz.* **4** (1964), 405–425.
9. G. M. VIANIKKO, *Amer. Math. Soc. Transl.* **86** (1970), 249–259.
10. M. A. KRASNESOL'SKII AND G. M. VAINIKKO, *et al.*, "Approximate solution of Operator Equations," Nauka, Moscow (1969) (Russian).
11. G. FIX, "On the approximation of eigenvalues arising from non-self-adjoint problems," University of Maryland Report, 1972.
12. I. BABUSKA, "The Mathematical Foundations of the Finite Element Method," Academic Press, New York, to appear.
13. J. L. LIONS AND E. MAGENES, "Problèmes aux limites non homogènes et applications," Dunod, Paris, 1968.
14. B. KELLOGG, "On the Poisson equation with intersecting interfaces," Technical Note BN-643, University of Maryland, 1970.
15. I. BABUSKA, *Numer. Math.* **16** (1971), 322–333.
16. G. J. HABETLER AND M. A. MARTINO, "Existence theorems and spectral theory for the multigroup diffusion model," Proc. of Symposia in Applied Math., 11, A.M.S., Providence (1961).
17. K. YOSIDA, "Functional Analysis," Springer–Verlag, New York, 1965.
18. I. BABUSKA, "The finite element method with Lagrangian multipliers," University of Maryland Report BN-724 (1972).
19. G. FIX, S. GULATI, AND G. I. WAHOFF, On the use of singular functions with the finite element method, *J. of Comp. Physics* (1972), to appear.
20. N. DUNFORD AND J. SCHWARTZ, "Linear Operations," Vol. II, Interscience, New York, 1963.
21. T. KATO, "Perturbation Theory of Linear Operators," Springer–Verlag, New York, 1966.
22. J. H. BRAMBLE AND J. E. OSBORN, Univ. of Wisconsin M.R.C. Report 1232, June 1972.